Alternative Tests for Correct Specification of Conditional Predictive Densities

Barbara Rossi* and Tatevik Sekhposyan†

January 27, 2014

Abstract

We propose new methods for evaluating predictive densities that focus on the models' actual predictive ability in finite samples. The tests offer a simple way of evaluating the correct specification of predictive densities, either parametric or non-parametric. The results indicate that our tests are well sized and have good power in detecting mis-specification in predictive densities. An empirical application to the Survey of Professional Forecasters and a baseline Dynamic Stochastic General Equilibrium model shows the usefulness of our methodology.

Keywords: Predictive Density, Dynamic Mis-specification, Forecast Evaluation

Acknowledgments: We thank T. Clark, F. Diebold, G. Ganics, A. Inoue, A. Patton, B. Perron, F. Ravazzolo, N. Swanson, M. Watson, participants of the seminars at the CORE Louvain-la-Neuve, UNSW, Monash University, the 2012 Australasian Meetings of the Econometric Society, the 2012 Time Series Econometrics Workshop in Zaragoza, the 2013 CIREQ Time Series and Financial Econometrics Conference, the 2013 Applied Time Series Econometrics Workshop of St. Louis Fed, the 2013 UCL conference on Frontiers in Macroeconometrics, the 2013 Conference on Forecasting Structure and Time Varying Parameter Patterns in Rotterdam for comments. The views expressed in this paper are solely those of the authors and should not be attributed to the Bank of Canada.

J.E.L. Codes: C22, C52, C53

*ICREA-UPF, Barcelona GSE, and CREI. Carrer Ramon Trias Fargas, 25-27, Mercè Rodoreda bldg., 08005 Barcelona, Spain. E-mail: barbara.rossi@upf.edu

†Bank of Canada, 234 Laurier Avenue West, Ottawa, ON, K1A 0G9, Canada. E-mail: tsekhposyan@bankofcanada.ca

1 Introduction

Policy institutions are becoming interested in complementing point forecasts with an accurate description of uncertainty. For instance, they are interested not only in knowing whether inflation is below its target, but also in understanding whether the realized inflation rate was forecasted to be a low probability event ex-ante. In fact, if researchers underestimate the uncertainty around point forecasts, it is possible that an event with a fairly high likelihood of occurrence is forecasted to be a very low probability event. An accurate description of uncertainty is therefore important in the decision making process of economic agents and policymakers. The objective of this paper is to provide reliable tools for evaluating whether the uncertainty around point forecasts is correctly specified.

Many central banks periodically report fan charts to evaluate and communicate the uncertainty around point forecasts (e.g., see the Bank of England 2013 Inflation Report Fan Charts or the Economic Bulletin by the Bank of Italy, 2012, p. 45). Fan charts provide percentiles of the forecast distribution for variables of interest. Typically, central banks' fan charts are the result of convoluted methodologies that involve a variety of models and subjective assessments, although fan charts can be based on specific models as well.[1] Figure 1 plots fan charts for US output growth (left panel) and the Federal Funds rate (right panel) based on a representative Dynamic Stochastic General Equilibrium (DSGE) model widely used in academia and policymaking (discussed in detail later on). The fan charts display DSGE-based forecasts made in 2000:IV for the next four quarters. The shaded areas in the figures depict the deciles of the forecast distribution and provide a visual impression of the uncertainty around the point forecasts (in this case, the median, marked by a dashed line). Over the four quarterly horizons, uncertainty about output growth and interest rate forecasts has a very different pattern: the uncertainty surrounding output growth forecasts is constant across horizons, while for interest rates it depends on the horizon. The dark line in the figures plots the actual realized value of the target variable. Clearly, forecasts of interest rates were very imprecise (the realization is outside every forecast decile except for one-quarter-ahead horizon), whereas the DSGE model predicts output growth more accurately. In order to evaluate the DSGE-based forecast distributions, it is important to understand whether it is the description of uncertainty that was inaccurate or the realized values were indeed low probability events.

INSERT FIGURE 1 HERE

[1] See for instance Clements (2004) for a discussion on the Bank of England fan charts.

Currently available methodologies test whether the empirical distribution belongs to a given parametric density family with parameters evaluated at their pseudo-true values. Our paper derives new tools to evaluate whether predictive densities are correctly specified by focusing on evaluating their actual forecasting ability in the finite samples typically available to researchers. In other words, we test whether the predictive densities are correctly specified to match the parametric model and its estimation technique. Furthermore, our proposed tests can be used to make inference on non-parametric predictive densities. Importantly, our tests do not require the model to be dynamically correctly specified nor its disturbances to be serially uncorrelated; however, for completeness, we also discuss a version of the tests that hold when the model is dynamically correctly specified.

The advantage of this approach relative to the existing literature (e.g., reviewed in Corradi and Swanson, 2006b) is that it allows the researcher to evaluate whether the density forecast is correctly specified at the actual parameter estimates. In contrast, most of the literature focuses on testing the correct specification of predictive densities evaluated at the pseudo-true parameter values, which may not be representative of the models' actual forecasting ability in finite samples. We propose an approach where parameter estimation error is maintained under the null hypothesis, as in Amisano and Giacomini (2007). However, our approach is very different from Amisano and Giacomini (2007): the latter focus on model selection by comparing the *relative performance* of competing models' predictive densities, whereas we focus on evaluating the *absolute performance* of a model's predictive density.

Maintaining parameter estimation error under the null hypothesis has two advantages: (i) there is no need to correct the test statistics for parameter estimation error, since that is maintained under the null hypothesis; and (ii) the asymptotic distribution of the test statistics at the one-step-ahead horizon is nuisance parameter free and the critical values can be tabulated when the model is dynamically correctly specified. We derive our tests within a class of Kolmogorov-Smirnov and Cramér-von Mises-type tests commonly used in the literature and show that all our proposed tests have good size properties in small samples.

When mis-specification of the predictive density is detected, an important step is to understand the source of the mis-specification. Many tests that exist in the literature concentrate on testing for correct specification of predictive densities by testing a joint hypothesis of uniformity and independence. Lack of uniformity implies an incorrect unconditional probability (on average) that the actual realizations of the target variable match the model's predictive density. Lack of independence refers to a situation where, even if on average realizations are compatible with the model's predictive density (i.e., the unconditional prob-

ability is correct), the pattern of the rejections is non-random. Thus, the rejection of the correct specification could be driven either by the lack of uniformity or independence, and it is important to identify the source. To uncover the source of the mis-specification, we: (i) propose new tests of uniformity robust to violations of independence; (ii) discuss some tests of serial correlation robust to violations of uniformity that are available in the literature and could be used. The tests can be applied to either one-step-ahead or multiple-step-ahead predictive densities.

Our paper is related to a series of contributions which test whether observed forecasts could have been generated by a given theoretical model's distribution. Diebold et al. (1998, 1999) introduced the probability integral transform (PIT) into economics as a tool to test whether the empirical predictive distribution of surveys or empirical models matches the true, unobserved distribution that generates the data. Their approach tests for properties of the PITs, such as independence and uniformity, by treating the forecasts as primitive data, that is without correcting for estimation uncertainty associated with those forecasts.[2] Additional approaches proposed in the literature for assessing the correct calibration of predictive densities are the non-parametric approach by Hong and Li (2005) and the bootstrap introduced by Corradi and Swanson (2006 a,b,c).[3] The null hypothesis in Hong and Li (2005) and Corradi and Swanson (2006 a,b,c) is that of correct specification of the density forecast at the pseudo-true (limiting) parameter values. Although this framework enables predictive density evaluation when the models are dynamically mis-specified, it does not necessarily capture the actual measure of predictive ability that researchers are interested in, as in small samples the pseudo-true parameter values may not be representative of the actual marginal predictive ability of the regressors. In the approach we propose, the main test statistic is the same as Corradi and Swanson's (2006a) one, although the null hypothesis is very different: it targets evaluating density forecasts at the estimated parameter values (as opposed to their population values).

We provide empirical applications of our proposed tests to the density forecasts in the Survey of Professional Forecasters (SPF) as well as those produced by a baseline DSGE

[2] González-Rivera and Sun (2013) operate in a framework similar to Diebold et al. (1998, 1999) and test for correct specification utilizing PIT based autocontours. Their approach takes into account parameter estimation error and could be extended to access the proper calibration of multivariate densities.

[3] Corradi and Swanson (2006 a,c) discuss methods for in-sample evaluation of predictive densities, while Corradi and Swanson (2006b) generalize the tests to be applicable in out-of-sample. Hong, Li and Zhao (2007) provide with an out-of-sample counterpart of the Hong and Li (2005) in-sample tests. See also Bontemps and Meddahi (2012) for in-sample tests of distributional assumptions.

model. Note that the SPF is a particularly suitable application to demonstrate the usefulness of our methodologies, since it involves a situation where the models are unknown to practitioners, and thus it is impossible to specifically account for parameter estimation uncertainty. Our test uncovers that SPF density forecasts of both output growth and inflation are mis-specified. We find that DSGE model-based predictive densities are mis-specified, at least given one of the tests we consider.

The remainder of the paper is organized as follows. Section 2 introduces the notation and definitions. Section 3 presents results for tests of correct specification of density forecasts robust to dynamic mis-specification and Section 4 discusses issues related to the practical applicability of our test. In Section 5, we provide Monte Carlo evidence on the performance of our tests in small samples. Section 6 analyzes the empirical applications to SPF and DSGE density forecasts and Section 7 concludes.

2 Notation and Definitions

We first introduce the notation and discuss the assumptions about the data, the models and the estimation procedure. Consider a stochastic process $\{Z_t : \Omega \to R^{k+1}\}_{t=1}^{T}$ defined on a complete probability space (Ω, F, P). The observed vector Z_t is partitioned as $Z_t = (y_t, X_t')'$, where $y_t : \Omega \to R$ is the variable of interest and $X_t : \Omega \to R^k$ is a vector of predictors. We are interested in the true but unknown $h-$step-ahead conditional predictive density for the scalar variable y_{t+h} based on $F_t = \sigma(Z_1', ..., Z_t')'$, which is the true information set available at time t. We denote the density by $\phi_0(.)$.[4]

We assume that the researcher has divided the available sample of size $T + h$ into an in-sample portion of size R and an out-of-sample portion of size P, and obtained a sequence of h-step-ahead out-of-sample density forecasts of the variable of interest y_t using the information set $\Im_t$, such that $R + P - 1 + h = T + h$ and $h < \infty$ and $\Im_t \subseteq \mathcal{F}_t$. Note that this implies that the researcher observes a subset of the true information set. We also let $\Im_{t-R+1}^{t}$ denote the truncated information sets based on information set used by the researcher available between time $(t - R + 1)$ and time t.

Let the sequence of P out-of-sample estimates of conditional predictive densities evaluated at the ex-post realizations be denoted by $\left\{\phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right)\right\}_{t=R}^{T}$. The dependence

[4] The true conditional forecast density may depend on the forecast horizon. To simplify notation, we omit this dependence without loss of generality given that the forecast horizon is fixed. Furthermore, we use the symbols $\phi_0(.)$ and $\phi_t(.)$ to denote generic distributions and not necessarily a normal distribution.

on the information set is a result of the assumptions we impose on the in-sample parameter estimates, $\widehat{\theta}_{t,R}$. We require the parameters to be either re-estimated at each $t = R, ..., T$ over a window of R data including data indexed $t - R + 1, ..., t$ (rolling scheme) or be estimated only once using a sample including data indexed $1, ..., R$ (fixed scheme). In this paper we are concerned with direct multi-step forecasting, where the predictors are lagged h periods. In addition to being parametric (such as a normal distribution), the distribution $\phi_{t+h}(.)$ can also be non-parametric (as in one of the empirical applications in this paper).

Consider the probability integral transform (PIT), which is the cumulative density function (CDF) corresponding to $\phi_{t+h}(.)$ evaluated at the realized value y_{t+h}:

$$z_{t+h} = \int_{-\infty}^{y_{t+h}} \phi_{t+h}\left(u|\Im_{t-R+1}^{t},\right) du \equiv \Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right).$$

Let

$$\xi_{t+h}(r) \equiv \left(1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \leq r\right\} - r\right),$$

where $1\{.\}$ is the indicator function and $r \in [0,1]$. Consider $\Psi(r) = \Pr\{z_{t+h} \leq r\} - r$ and its out-of-sample counterpart:

$$\Psi_P(r) \equiv P^{-1/2} \sum_{t=R}^{T} \xi_{t+h}(r). \tag{1}$$

Let us also denote the empirical probability distribution function of the PIT by

$$\varphi_P(r) \equiv P^{-1} \sum_{t=R}^{T} 1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \leq r\right\}. \tag{2}$$

3 Asymptotic Tests of Specification

This section presents results for the case of one-step-ahead forecasts when the densities are dynamically correctly specified; we then generalize the tests to the presence of mis-specification and serial correlation. The generalized case could also apply to the $h > 1$ step-ahead forecasts. All the proofs are relegated to the Appendix A. The tests we propose have an asymptotic distribution that is free of nuisance parameters in the one-step-ahead forecast case when the models are dynamically correctly specified. In this case the critical values can be tabulated. We also discuss tests that are valid for multi-step-ahead forecasts and in the presence of dynamic mis-specification. Both of these cases introduce serial correlation in the dynamics of the PITs.

In order to maintain parameter estimation error under the null hypothesis, we state our null hypothesis in terms of a truncated information set, which expresses the dependence of the predictive density on estimated parameter values (as in Amisano and Giacomini, 2007). We focus on testing $\phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right)=\phi_0\left(y_{t+h}|\mathcal{F}_t\right)$, that is:

$$H_0:\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right)=\Phi_0\left(y_{t+h}|\mathcal{F}_t\right), \tag{3}$$

where $\Phi_0\left(y_{t+h}|\mathcal{F}_t\right)\equiv\Pr\left(y_{t+h}\leq y|\mathcal{F}_t\right)$ denotes the distribution specified under the null hypothesis. The alternative hypothesis, H_A, is the negation of H_0. Note that the null hypothesis evaluates the correct specification of the density forecast of a model estimated with a given window size, R, as well as the parameter estimation method chosen by the researcher.

We are interested in the test statistics:

$$\kappa_P^{CS}=\sup_{r\in[0,1]}\Psi_P\left(r\right)^2, \tag{4}$$

$$C_P^{CS}=\int_0^1\Psi_P\left(r\right)^2dr. \tag{5}$$

Note that the κ_P^{CS} test statistic is basically the same as the V_{1T} test statistic considered by Corradi and Swanson (2006a) when applied to predictive densities (the latter consider the absolute value of $\Psi_P\left(r\right)$, while we consider its square). Note, however, that we derive the asymptotic distribution of the test statistic under a different null hypothesis. Corradi and Swanson (2006a) focus on the null hypothesis: $H_0^{CS}:\Phi_{t+h}(y_{t+h}|\Im_t)=\Phi_0(y_{t+h}|\Im_t,\theta^\dagger)$ for some $\theta^\dagger\in\Theta$, where Θ is the parameter space. That is, Corradi and Swanson (2006a) test the hypothesis of correct specification of the predictive density at the pseudo-true parameter value. Thus, the limiting distribution of their test reflects parameter estimation error and, therefore, is not nuisance parameter free. In addition, they allow for dynamic mis-specification under the null hypothesis. This allows them to obtain asymptotically valid critical values even when the information set may not contain all the relevant past history. Dynamic mis-specification also affects the limiting distribution of their test statistic by contributing additional nuisance parameters.

Under our null hypothesis in eq. (3) instead, the limiting distribution of the test statistic is nuisance parameter free when the model is dynamically correctly specified. The reason is that we maintain parameter estimation error under the null hypothesis, which implies that the asymptotic distribution of the test does not require a delta-method approximation around the pseudo-true parameter value.

To clarify our null hypothesis and understand how it differs from the null hypothesis considered in traditional tests, we provide a couple of examples.

Example 1: *As a simple example, consider the case where the true data are generated by* $y_t = \theta + \eta_t$ *where* $\theta = 1$ *(without loss of generality), and* $\eta_t \sim iid\ N(0,1)$*, where iid denotes independent and identically distributed. The (conditional) distribution of* y_t *is normal, with mean* 1 *and variance 1.*

Suppose the forecaster uses the correctly specified model and estimates the parameter $\widehat{\theta}_{t,R}$ *using a window of size* $R = 2$ *while* σ^2 *is the in-sample variance of the errors, which is assumed to be known. Then,*

$$\widehat{\theta}_{t,2} = \frac{1}{2}(y_t + y_{t-1}) \tag{6}$$

and the predictive density used by the forecaster is

$$\phi_{t+1}\left(u|\Im_{t-R+1}^{t}\right) = (2\pi\sigma^2)^{-1/2} \exp\left(-\tfrac{1}{2}\frac{\left(u-\frac{1}{2}(y_t+y_{t-1})\right)^2}{\sigma^2}\right).$$

Note that the resulting predictive distribution is different from the distribution of the data generating process, since one would never consistently estimate the unknown parameter with only two observations. However, our objective is to test whether the distribution is correctly specified given the forecasting model and its estimation technique, not to test the correct specification relative to the data generating process. In other words, we are not interested in the distribution of y_{t+1} *given the true parameter value (even when we are using the correctly-specified parametric model for forecasting), which is* $N(1,1)$*. Rather, we are interested in testing whether the forecasts, conditional on the forecasting information set, are normally distributed. That is, whether* $\phi_{t+1}\left(u|\Im_{t-R+1}^{t}\right)$ *is normally distributed with the mean and the variance specified above. Since* $\phi_0(u|\mathcal{F}_t) = N(1,1)$ *and* $\phi_{t+1}\left(u|\Im_{t-R+1}^{t}\right) = N\left(\frac{1}{2}(y_t+y_{t-1}), \sigma^2\right) \neq \phi_0(u|\mathcal{F}_t)$*, the null hypothesis typically considered in the literature (e.g. Corradi and Swanson, 2006b) does hold, although our null hypothesis does not hold.*

Conversely, let the true data generating process be $y_{t+1} = \frac{1}{2}(y_t + y_{t-1}) + \sigma\eta_t$*, with a conditional distribution of* $\phi_0(u|\mathcal{F}_t) = N\left(\frac{1}{2}(y_t+y_{t-1}), \sigma^2\right)$*. Yet the researcher obtains its forecasts using a mis-specified model with a constant and* $R = 2$*. Note that, in this case,* $\Im_{t-R+1}^{t} = \mathcal{F}_t$*. Then,* $\phi_{t+1}\left(u|\Im_{t-R+1}^{t}\right) = (2\pi\sigma^2)^{-1/2} \exp\left(-\tfrac{1}{2}\frac{\left(u-\frac{1}{2}(y_t+y_{t-1})\right)^2}{\sigma^2}\right)$*, which is the same distribution as the one generating the data,* $\phi_0(u|\mathcal{F}_t)$*, and our null hypothesis holds notwithstanding the fact that the researcher is estimating a mis-specified model.*

Therefore, the null hypothesis in eq. (3) does not test the correct specification of the forecast model evaluated at the true parameter values relative to the data generating process;

rather, the null hypothesis in eq. (3) tests the correct specification of the forecast model evaluated at the parameter values obtained conditional on the estimation procedure. We argue that the latter is the appropriate approach to evaluate the correct specification of density forecasts.

While the example is provided to help with intuition and thus is very simple, it might be unrealistic. Note, however, that the same broad conclusions would hold in more general cases; for instance, when estimating the rolling parameters with weighted averages, where $\widehat{\theta}_{t,R} = \sum_{j=t-R+1}^{t} \omega_j y_j$ *rather than the simple case in eq. (6). Note also that the example can be generalized to a wider set of models other than the one with only a constant. In the case of a simple autoregressive (AR) model of order 1, the conditional mean at time t is:* $\left(\sum_{j=t-R+1}^{t} y_{j-1}^2\right)^{-1} \left(\sum_{j=t-R+1}^{t} y_{j-1} y_j\right) y_t$. *The methodology requires only that the conditional mean be estimated based on a finite number of observations.*[5]

Example 2: *Suppose the true data generating process is:* $y_t = \alpha + x_t + \varepsilon_t$ *where* $x_t \sim iid\ \chi_1^2$ *and* $\varepsilon_t \sim iid\ N(0,1)$. *Let the researcher estimate the mis-specified model that has only a constant (and does not include* x_t*). Thus, the (conditional) mean for* y_{t+1} *estimated at time t is:* $\widehat{\theta}_{t,R} \equiv \frac{1}{R} \sum_{j=t-R+1}^{t} y_j = \alpha + \frac{1}{R} \sum_{j=t-R+1}^{t} x_j + \frac{1}{R} \sum_{j=t-R+1}^{t} \varepsilon_j$. *The goal of the researcher is to test whether the forecast distribution is normally distributed. Note that the forecast will not be normally distributed (due to the mis-specification, the forecast distribution* $\phi_{t+1}\left(u|\Im_{t-R+1}^{t}\right)$ *will be a mixture of normals and chi-squares) and thus the null hypothesis does not hold even if the data is generated under the normality assumption. Again, the reason is that our focus is on testing the correct specification of the forecast density, not whether the true data generating process is normally distributed.*

3.1 One-step-ahead Density Forecasts and Dynamically Correctly Specified Models

This sub-section presents results for the case of one-step-ahead forecasts when the densities are dynamically correctly specified; the next section generalizes the tests to the presence of mis-specification and serial correlation. Let $h = 1$. First, we derive the asymptotic

[5] *The results in this paper also carry over to the fixed-estimation scheme, where the conditioning information set is* $\Im_1^R$*, or to any other information set based on a bounded number of observations R, provided R is finite.*

distribution of $\Psi_P(r)$ for one-step-ahead density forecasts under Assumption 1.

Assumption 1.

(i) $\left\{Z_t = (y_t, X_t')'\right\}_{t=R}^{T}$ is mixing with $\phi(j)$ of size $-\lambda/(2\lambda-1)$ when $\lambda \geq 1$ or $\alpha(j)$ of size $-\lambda/(\lambda-1)$ when $\lambda > 1$. y_{t+1} is generated from $\{\phi_0(y_{t+1}|F_t)\}_{t=R}^{T}$ with a cumulative distribution function $\Phi_0(.)$ that is continuous, differentiable and has a well defined inverse;

(ii) $\left\{\Phi_{t+1}^{-1}\left(z_{t+1}|\Im_{t-R+1}^{t}\right)\right\}_{t=R}^{T}$ has non-zero Jacobian with continuous partial derivatives;

(iii) $R < \infty$ as $P, T \to \infty$.

In this section, we focus on dynamically correctly specified models. The more general case of dynamic misspecification will be discussed in the next section. Dynamic correct specification is characterized by Assumption 2:

Assumption 2.

$y_{t+h}|\Im_{t-R+1}^{t} \equiv y_{t+h}|\mathcal{F}_t$ for all $t = R, ..., T$, where $\equiv$ denotes equality in distribution.

Theorem 1 (Asymptotic Distribution of $\Psi_P(r)$) *Under Assumptions 1, 2, and H_0 in eq. (3): (i) $\{z_{t+1}\}_{t=R}^{T}$ is iid $U(0,1)$; (ii) $\Psi_P(r)$ weakly converges as a variable in the space $[0,1] \times \mathbb{R}$) to the Gaussian process $\Psi(.)$, with mean zero and auto-covariance function $E[\Psi(r_1)\Psi(r_2)] = [\inf(r_1, r_2) - r_1 r_2]$.*

The result in Theorem 1 allows us to derive the asymptotic distribution of the test statistics of interest, presented in Theorem 2. The latter shows that the asymptotic distribution of our proposed test statistics have the appealing feature of being nuisance parameter free.

Theorem 2 (Correct Specification Tests) *Under Assumptions 1, 2 and H_0 in eq. (3):*

$$\kappa_P^{CS} \equiv \sup_{r\in[0,1]} \Psi_P(r)^2 \Rightarrow \sup_{r\in[0,1]} \Psi(r)^2, \tag{7}$$

and

$$C_P^{CS} \equiv \int_0^1 \Psi_P(r)^2 dr \Rightarrow \int_0^1 \Psi(r)^2 dr. \tag{8}$$

Reject H_0 at the $\alpha \cdot 100\%$ significance level if $\kappa_P^{CS} > \kappa_\alpha^{CS}$ and $C_P^{CS} > C_\alpha^{CS}$. Critical values for $\alpha = 10\%$, 5% and 1% are provided in Table 1, Panel A.

INSERT TABLE 1 HERE

Note that one could be interested in testing correct specification in specific parts of the distribution.[6] For example, one might be interested in the tails of the distribution, which correspond to outliers, such as the left tail where $r \in (0, 0.25)$, or the right tail where $r \in (0.75, 1)$, or both: $r \in \{(0, 0.25 \cup 0.75, 1)\}$. Alternatively, one might be interested in the central part of the distribution, for example $r \in [0.25, 0.75]$. We provide critical values for these interesting cases in Table 1, Panel B.

Note also that our κ_P^{CS} test has a graphical interpretation. In fact,

$$\alpha = \Pr\left\{\sup_{r\in[0,1]} \Psi_P(r)^2 > \kappa_\alpha^{CS}\right\} = \Pr\left\{\left[\sup_{r\in[0,1]} |\Psi_P(r)|\right]^2 > \kappa_\alpha^{CS}\right\} = \Pr\left\{\sup_{r\in[0,1]} |\Psi_P(r)| > \sqrt{\kappa_\alpha^{CS}}\right\}.$$

Thus, from eqs. (1) and (2),

$$\frac{1}{\sqrt{P}}\Psi_P(r) \equiv P^{-1}\sum_{t=R}^{T}\left(1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \leq r\right\} - r\right) = \varphi_P(r) - r.$$

Furthermore,

$$\alpha = \Pr\left\{\sup_{r\in[0,1]} |\Psi_P(r)| > \sqrt{\kappa_\alpha^{CS}}\right\} = \Pr\left\{\sup_{r\in[0,1]} |\varphi_P(r) - r| > \sqrt{\kappa_\alpha^{CS}/P}\right\}.$$

This suggests the following implementation: plot the cumulative distribution function of the PIT, eq. (2), together with the cumulative distribution function of the uniform, r (the 45-degree line), and the critical value lines: $r \pm \sqrt{\kappa_\alpha^{CS}/P}$. Then, the κ_P^{CS} test rejects if the cumulative distribution function of the PIT is outside the critical value lines. It also follows from this argument that the critical values of the test statistic $\sup_{r\in[0,1]} |\Psi_P(r)|$ would be $\sqrt{\kappa_\alpha^{CS}}$.

It is interesting to compare our approach to Diebold et al. (1998). While our null hypothesis is different from theirs, the procedure that we end up proposing is very similar to theirs in that both their implementation and ours abstract from parameter estimation error. Thus, our approach can be viewed as a formalization of the approach suggested in Diebold et al. (1998), albeit with a different null hypothesis. An additional advantage of our approach is that the confidence bands that we propose are joint, not pointwise.

The previous discussion suggests that we could also apply our approach to likelihood-ratio (LR) tests based on the inverse normal transformation of the PITs. It is well known that, when the forecast density is correctly specified, an inverse normal transformation of the

[6] See Franses and van Dijk (2003), Amisano and Giacomini (2007) and Diks, Panchenkob and van Dijk (2011) for a similar idea in the context of point forecasts and density forecast comparisons.

PITs (ζ_{t+h}) has a standard normal distribution (Berkowitz, 2001). As noted in the literature, the latter approach has typically abstracted from parameter estimation uncertainty. When focusing on the traditional null hypothesis H_0^{CS}, ignoring parameter estimation error leads to size distortions. Note that the size distortion is not only a small sample phenomenon, but persists asymptotically. The next result shows that, since parameter estimation error is maintained under our null hypothesis H_0, eq. (3), there is no need to correct the asymptotic distribution and the implied critical values of the likelihood ratio tests to account for parameter estimation error.

Corollary 3 (Inverse Normal Tests) *Let $\Phi^{-1}(.)$ denote the inverse of the standard normal distribution function. Under Assumptions 1,2 and H_0 in eq. (3): $\zeta_{t+1} \equiv \Phi^{-1}(z_{t+1})$ is iid $N(0,1)$.*

Thus, one could test for the correct specification of the density forecast by testing the absence of serial correlation and the correct specification of the moments of ζ_{t+h}. For example, estimate an AR(1) model for ζ_{t+1} and test that the mean and the slope are both zero, and that the variance is one. The advantage of this approach is that it is informative regarding the possible causes underlying the mis-specification of the density forecast and it may perform better in small samples. The disadvantage of the approach is that, unlike the κ_P^{CS} and C_P^{CS} tests, it focuses on specific moments of the distribution rather than the whole (non-parametric) cumulative distribution function.

Also, our approach can be generalized to predictive scores, although in this case the asymptotic distribution depends on nuisance parameters. We provide details in the next sub-section.

Finally, note that our approach provides not only a rationale to the common practice of evaluating the correct specification of density forecasts using PITs without adjusting for parameter estimation error (Diebold et al., 1998), but also a methodology for implementing tests robust to the presence of serial correlation as well as dynamically mis-specified models. This is a more general case and we consider it in the next section.

3.2 Multi-step-ahead Forecasts and Dynamic Mis-specification

When considering h-step-ahead forecasts, $h > 1$ and finite, as well as when $h = 1$ for models that are dynamically mis-specified, an additional problem arises, as both of these cases involve serial correlation in the PITs.[7] Thus, we need to extend our results and

[7] In fact, h-step-ahead forecasts are serially correlated of order at least $(h-1)$.

allow the forecasts to be both serially correlated and potentially mis-specified under the null hypothesis. Consider the following Assumption:

Assumption 3.

(i) $\left\{Z_t = (y_t, X_t')'\right\}_{t=R}^{T}$ is strong mixing with $\alpha(j)$ of size $-\lambda/(\lambda-1)$, where $\lambda \in [1,2)$, $\sum_{j=1}^{\infty} j^2 \alpha(j)^{\lambda/(4+\lambda)} < \infty$. y_{t+h} is generated from $\{\phi_0(y_{t+h}|\mathcal{F}_t)\}_{t=R}^{T}$, whose cumulative distribution function $\Phi_0(.)$ is continuous, differentiable and has a well defined inverse;

(ii) $\Pr\left(\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \leq r_1, \Phi_{t+h+d}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \leq r_2\right) = F_d(r_1, r_2)$,

$\Pr\left(\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \leq r\right) = F(r)$, where $F_d(.,.)$ and $F(.)$ are the distribution functions of the random variable $\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right)$ and $F(.)$ is continuous.

Under serial correlation or h-step-ahead forecasts, we show that $\Psi_P(r)$ weakly converges (considered as variables in the space $[0,1]\times\mathbb{R}$) to the Gaussian process $\Psi(.,.)$, with mean zero and an auto-covariance function that depends on the serial correlation.

Theorem 4 (Correct Specification Tests under Serial Correlation) *Under Assumptions 1(ii), 1(iii), 3 and H_0 in eq. (3), $\Psi_P(r)$ weakly converges considered as a variable in the space $[0,1]\times\mathbb{R}$) to the Gaussian process $\Psi(.)$, with mean zero and auto-covariance function $E[\Psi(r_1)\Psi(r_2)] = \sigma(r_1, r_2)$, where $\sigma(r_1, r_2) = \sum_{d=-\infty}^{\infty}[F_d(r_1, r_2) - F(r_1)F(r_2)]$. Furthermore,*

$$\kappa_P^{CS} \Rightarrow \sup_{r\in[0,1]} \Psi(r)^2,$$

$$C_P^{CS} \Rightarrow \int_0^1 \Psi(r)^2 dr.$$

For a given estimate of $\sigma(r_1, r_2)$, the critical values of κ_P^{CS} and C_P^{CS} can be obtained via Monte Carlo simulations.[8]

However, there are several other solutions proposed in the literature that one could use within our approach as well. A first approach is to discard data by reducing the effective sampling rate to ensure an uncorrelated sample (Persson, 1974 and Weiss, 1973). This can be implemented in practice when models are dynamically correctly specified by creating subsamples of forecast distributions that are at least h periods apart. However, this procedure may not be possible in small samples, since sub-sampling may significantly reduce the size of the sample. In those cases, one may implement the procedure in several uncorrelated sub-samples of forecasts that are at least h periods apart and then use Bonferroni methods

[8] In the Monte Carlo and in the empirical application we used Newey and West's (1987) HAC estimator to calculate the covariance of the PITs.

to obtain a joint test without discarding observations (see Diebold et al., 1998). However, it is well-known that Bonferroni methods are conservative; thus the latter procedure, while easy to implement, may suffer from low power. Alternative approaches include using a block bootstrap to obtain the critical values (Bai and Ng, 2005, and Corradi and Swanson, 2006c). The relative merit and validity of these approaches remains to be studied further.

It is interesting to note that our approach can be applied to predictive scores. We consider the case of the Continuous Rank Probability Score (CRPS, Gneiting and Raftery, 2007) in Appendix A. The asymptotic distribution of the CRPS is not nuisance parameter free even when predictive densities are correctly specified and $h = 1$. However, the critical values can be obtained using Monte Carlo simulations.

4 How to Use Our Tests?

Suppose the researcher decides to use the tests described in Theorem 2. If the tests reject, the rejection could be due to the violation of either independence or uniformity. If the researcher is concerned about the failure of independence, he/she could use our test for uniformity robust to violations of independence. As discussed in Section 3.2, the latter test is not nuisance parameter free, so the implementation is more challenging and requires simulating the critical values. Alternatively, one could test for serial correlation in a way that is robust to uniformity. Among the tests that could be implemented, one could consider the Ljung-Box Q or Box-Pierce Q-test statistics (Box and Pierce, 1970) or the BDS test proposed by Brock, Dechert and Scheinkman (1987). The Q-test detects auto-correlation in a linear framework whereas the BDS test is a non-parametric test of independence and identical distribution against an unspecified alternative. Note that serial correlation implies lack of independence but serial uncorrelatedness does not necessarily imply independence. If the PITs pass the independence test, the researcher should feel more comfortable in applying the critical values provided in the paper.

5 Monte Carlo Evidence

In this section we analyze the size and power properties of our proposed tests in small samples for both correctly specified and mis-specified forecasting models.

5.1 Size Analysis

To investigate the size properties of our tests we consider several Data Generating Processes (DGPs). The forecasts are based on model parameters estimated in rolling windows for $t = R, ..., T+h$. We consider several values of $R = [50, 100, 200]$ and $P = [100, 200, 500, 1000]$ to evaluate the performance of the proposed procedure in finite samples. While our Assumptions require R finite, we investigate both small and large values of R to investigate how robust the performance of the methodology is when R is large. The construction of the DGPs that ensure that the null hypothesis holds is inspired by Amisano and Giacomini (2007, p. 181).[9] The DGPs are the following:

DGP S1 (Baseline Model): We estimate a model with a constant. To ensure that the null hypothesis in eq. (3) holds**,** we generate the data under the null hypothesis according to $\widetilde{y}_{t+1} = R^{-1} \sum_{j=t-R+1}^{t} y_j + \widehat{\sigma}_t \eta_{t+1}$, $t = R, ..., T$, $y_j = \mu + \varepsilon_j$, $\varepsilon_j \sim iid\ N(0,1)$, $\mu = 5$ and $\eta_t \sim iid$ $N(0,1)$ independent of ε_t, where $\widehat{\sigma}_t^2 = R^{-1} \sum_{j=t-R+1}^{t} (y_j - R^{-1} \sum_{i=t-R+1}^{t} y_i)^2$.[10]

DGP S2 (Estimated Model): We parameterize the model according to the realistic situation where the researcher is interested in forecasting one-quarter-ahead real GDP growth with a bivariate model that includes both lagged GDP growth as well as lagged term spread in U.S. data from 1959:I-2010:III. The forecasts are generated under the null hypothesis: $\widetilde{y}_{t+1} = Z_t' \left(R^{-1} \sum_{j=t-R+1}^{t} Z_j Z_j' \right)^{-1} \left(R^{-1} \sum_{j=t-R+1}^{t} Z_j y_j \right) + \widehat{\sigma}_t \eta_{t+1}$, where $Z_j = (1, y_j, y_{j-1}, x_j, x_{j-1})'$, $y_j = \mu + \beta_1 y_{j-1} + \beta_2 y_{j-2} + \gamma_1 x_{j-1} + \gamma_2 x_{j-2} + \varepsilon_j$, $\varepsilon_j \sim iid\ N(0, 9.54^2)$, $x_j = 0.2 + 0.8 x_{j-1} + \nu_j$, $\nu_j \sim iid\ N(0, 1.08^2)$ independent from ε_t, $\mu = 1.50$, $\beta_1 = 0.20$, $\beta_2 = -0.22$, $\gamma_1 = 0.13$, $\gamma_2 = 0.82$, and $\widehat{\sigma}_t^2 = R^{-1} \sum_{j=t-R+1}^{t} (y_j - (R^{-1} \sum_{i=t-R+1}^{t} Z_i Z_i')^{-1} (R^{-1} \sum_{k=t-R+1}^{t} Z_k y_k))^2$.[11]

DGPs S1-S2 are based on one-step-ahead forecast densities. DGP S3 considers the case of h-step-ahead forecast densities and serial correlation.

DGP S3 (Serial Correlation): The DGP is: $\widetilde{y}_t = R^{-1} \sum_{j=t-R+1}^{t} y_j + u_t + \rho u_{t-1}$, where

[9] In particular, note that the errors are drawn twice in order to satisfy the null hypothesis in a computationally feasible way.

[10] The results are unchanged if a different value of μ is considered.

[11] The lag length for y_t is selected by BIC in-sample: we first choose the best-fitting autoregressive model by BIC, then augment it with the optimal lags of the variable x_t selected again by BIC. The unconditional means of output growth and term spread are estimated from the data, and are 3.09 and 0.92 respectively. The process for the term spread (x_t) is an $AR(1)$.

$y_j = \mu + \varepsilon_j + \rho\varepsilon_{j-1}$, $\varepsilon_j \sim iid\ N(0,1)$, $\rho = 0.2$ and $u_j \sim iid\ N\left(0, \widehat{\sigma}_j^2\right)$, $\widehat{\sigma}_j^2 = R^{-1} \sum_{k=j-R+1}^{j} (y_k - R^{-1} \sum_{s=j-R+1}^{j} y_s)^2$.

The results are shown in Tables 2 and 3. Table 2 shows that our tests performs very well in finite samples, with only very mild under-rejections for small values of R and P for the Kolmogorov-Smirnov-type test. Table 3 shows that, in the case of serial correlation, the asymptotic distribution of the tests in Theorem 3 approximated using HAC-consistent variance estimates tends to over-reject in finite samples, although mildly.

INSERT TABLE 2 HERE

5.2 Power Analysis

To investigate the power properties of our tests, we consider the case of constant mis-specification in the following DGP.

DGP P: The data are generated from a linear combination of normal and χ_1^2 distributions: $\widetilde{y}_t = R^{-1} \sum_{j=t-R+1}^{t} y_j + (1-c)\,\widehat{\sigma}_t \eta_{1,t} + c\left(\eta_{2,t}^2 - 1\right)\sqrt{2}$, where $y_j = \mu + \varepsilon_j$, $\mu = 1$ and $\widehat{\sigma}_t^2 = R^{-1} \sum_{j=t-R+1}^{t} (y_j - R^{-1} \sum_{s=t-R+1}^{t} y_s)^2$. Furthermore, $\varepsilon_j, \eta_{1,t}$ and $\eta_{2,t}$ are $iid\ N(0,1)$ random variables that are independent of each other. The researcher tests whether the data result from a normal distribution, i.e. whether $\widetilde{y}_t \sim iid\ N\left(R^{-1} \sum_{j=t-R+1}^{t} y_j, \widehat{\sigma}_t\right)$. When c is zero, the null hypothesis is satisfied. When c is positive, the considered density becomes a convolution of a standard normal and a χ_1^2 distribution (with mean zero and variance one), where the weight on the latter becomes larger as c increases.[12]

The results shown in Table 4 suggest that our proposed specification tests $(\kappa_P^{CS}, C_P^{CS})$ have good power properties in detecting mis-specification in the predictive density.

INSERT TABLE 3 HERE

[12] Note that $\left(\eta_{2,t}^2 - 1\right)\sqrt{2}$ is a chi-squared distribution with zero mean and variance one, that is, it has the same mean and variance as the normal distribution we have under the null hypothesis, although the shape is different.

6 Empirical Analysis

This section provides an empirical assessment of the correct specification of widely-used density forecasts: the Survey of Professional Forecasters' (SPF) density forecasts of inflation and output growth, and density forecasts of the seven macroeconomic aggregates in Smets and Wouters' (2007) DSGE model.

6.1 Evaluation of SPF Density Forecasts

Diebold et al. (1999) evaluate the correct specification of the density forecasts of inflation in the SPF.[13] In this section, we conduct a formal test of correct specification for the SPF density forecasts using our proposed procedure and compare our results to theirs. In addition to inflation, we also investigate the conditional density forecasts of output growth. Note that the SPF only provides the forecasts of the survey participants, but not details on how the forecasts are obtained. Thus, it is impossible to know the true model that generated the forecasts. However, our methodology does not require the researcher to know that information, since it is subsumed in the null hypothesis in eq. 3.

We use real GNP/GDP and the GNP/GDP deflator as measures of output and prices. The mean probability distribution forecasts are obtained from the Federal Reserve Bank of Philadelphia. In the SPF data set, forecasters are asked to assign a probability value (over pre-defined intervals) of year-over-year inflation and output growth for the current (nowcast) and following (one-year-ahead) calendar years. The forecasters update the assigned probabilities for the nowcasts and the one-year-ahead forecasts on a quarterly basis. The probability distribution provided by the SPF is discrete, and we base our results on a continuous approximation by fitting a normal distribution. The realized values of inflation and output growth are based on the real-time data set for macroeconomists, also available from the Federal Reserve Bank of Philadelphia.[14]

The analysis of the SPF probability distribution is complicated since the SPF questionnaire has changed over time in various dimensions: there have been changes in the definition of the variables, the intervals over which probabilities have been assigned, as well as the

[13] SPF provides two types of density forecasts: one is the distribution of point forecasts across forecasters (which measures the dispersion of point forecasts across forecasters), and the other is the mean of the probability density forecasts (which measures the average of the density forecasts across forecasters). We focus on the latter.

[14] The data are available at http://www.philadelphiafed.org/research-and-data/real-time-center.

time horizon for which forecasts have been made. To mitigate the impact of these problematic issues, we truncate the data set and consider only the period 1981:III-2011:IV. We use the year-over-year growth rates of output and prices calculated from the first quarterly vintage of real GNP/GDP and the GNP/GDP deflator in each year to evaluate the density forecasts. For instance, in order to obtain the growth rate of real output for 1981, we take the 1982:I vintage of data and calculate the growth rate of the annual average GNP/GDP from 1980 to 1981. We consider the annual-average over annual-average percent change (as opposed to fourth-quarter over forth-quarter percent change) in output and prices to make it comparable with the definition of the variables that SPF forecasters provide probabilistic predictions for.

The empirical results are shown in Table 4. Asterisks ('*') indicate rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4. The latter are simulated conditional on the data (i.e. conditional on the HAC estimate of the variance-covariance matrix for the PIT). The tests which are robust to violations of independence under the null hypothesis (based on Theorem 4) consistently reject the null hypotheses of correct specification for both output growth and inflation. The non-robust test also rejects correct specification, except for output growth at the one-quarter-ahead forecast horizon.

INSERT TABLE 4 HERE

Our results are important in light of the finding that survey forecasts are reportedly providing the best forecasts of inflation. For example, Ang et al. (2007) find that survey forecasts outperform other forecasting methods (including the Phillips curve, the term structure and ARIMA models) and that, when combining forecasts, the data put the highest weight on survey information. Our results imply that survey forecasts still do not provide correct forecasts for the whole distribution of inflation.

Figure 2 plots the empirical CDF of the PITs (solid line). Under the null hypothesis in Theorem 2, the PITs should be uniformly distributed; thus the CDF of the PITs should be the 45 degree line. The figure also reports the critical values based on the κ_P^{CS} test. If the empirical CDF of the PITs is outside the critical value lines, we conclude that the density forecast is mis-specified. Clearly, the correct specification is rejected in all cases except the one-year-ahead density forecast of GDP growth. The figure also provides a visual analysis of

the mis-specification in the PITs: the survey typically overpredicts future large realizations (both positive and negative) of output growth and inflation.

For comparison, Figure 3 reports results based on Diebold et al.'s (1998) test. Panel A in Figure 3 plots the empirical distribution of the PITs of output growth for both the density nowcast (left-hand panel) and the one-year-ahead density forecast (right-hand panel). In addition to the PITs, we also provide the 95% confidence interval (dotted lines) using a normal approximation to a binomial distribution similar to Diebold et al.'s (1998). Both nowcast and one-year-ahead density forecasts of output growth are mis-specified, although mis-specification is milder in the case of one-year-ahead output growth. Figure 3, Panel B, shows instead the PITs for inflation. According to this test, both the density nowcast and one-year-ahead forecast overestimate tail risk. This phenomenon is more pronounced for the nowcast. Overall, the results obtained by using Diebold et al.'s (1998) test are broadly similar to those obtained by using the test that we propose in this paper, with one important exception. In the case of one-year-ahead GDP growth forecasts, our test based on Theorem 2 does not reject, whereas the Diebold et al. (1998) test does, despite the fact that both rely on *iid* assumptions. The discrepancy in the results is most likely due to the fact that the latter test is pointwise, whereas we jointly test the correct specification across all quantiles in the empirical distribution function: thus our test has larger critical values than the latter, in order to correctly account for the joint null hypothesis.

INSERT FIGURES 2 AND 3 HERE

6.2 Evaluation of a Baseline DSGE Model

DSGE models are widely used in central banks for policy evaluation and forecasting. Several recent contributions have focused on the ability of DSGE models to produce good out-of-sample point forecasts. In particular, Smets and Wouters (2007) show that the forecasts of the DSGE model that they propose are competitive relative to Bayesian VAR forecasts. Edge, Kiley and Laforte (2010) evaluate the predictive ability of the Federal Reserve Board's DSGE model (Edo), and Edge and Gürkaynak (2010) provide a thorough analysis of the forecasting ability of DSGE models using real-time data. The main result in the latter is that point forecasts of DSGE models perform similarly to that of a constant mean model, but both are biased; the reason why they perform similarly is because volatility was low during the Great Moderation sample period they consider, and, therefore, most variables were unpredictable. Edge, Gürkaynak and Kısacıkoğlu (2013) extend the results of Edge

and Gürkaynak (2010) to a longer sample and Gürkaynak, Kısacıkoğlu and Rossi (2013) analyze the point forecasting ability of the models relative to reduced-form models, and find that the latter perform better than the DSGE model at some forecast horizons.[15]

While the contributions discussed above focus on evaluating how accurate DSGE models' point forecasts are, central banks are becoming more and more interested in analyzing the uncertainty around the point forecasts that DSGE models provide. In this section, we focus on evaluating density forecasts of a baseline DSGE model. Only a few recent contributions analyze density forecast performance of DSGE models. Christoffel, Coenen and Warne (2010) study the performance of density forecasts of the European Central Bank's DSGE model (NAWM) and find that it tends to overestimate nominal wages. Wolters (2012) evaluates point and density forecasts of DSGE models for US inflation, output growth and the interest rate by comparing them to non-structural large dataset models and weighted forecasts. This paper also evaluates the accuracy of density forecasts and concludes that the DSGE models overestimate uncertainty around point forecasts. Bache, Jore, Mitchell and Vahey (2011) combine density forecasts of inflation from VARs and a DSGE model using the linear opinion pool. They find that allowing for structural breaks in the VAR produces well-calibrated density forecasts for inflation but reduces the weight on the DSGE considerably. Our paper differs from the literature as we evaluate DSGE model-based density forecasts; we do so by using our novel PIT-based test and compare its results with those based on the PIT-based tests proposed by Diebold et al. (1998).

We focus on the Smets and Wouters (2007) model as the benchmark DSGE model. Smets and Wouters' (2007) DSGE model is a real business cycle model with both nominal as well as real rigidities; in fact, it features sticky prices and wages as well as habit formation in consumption and cost of adjustment in investment.[16] We recursively re-estimate the Smets and Wouters (2007) model using exactly their data and priors in fixed rolling window of 80 observations and produce a sequence of 80 out-of-sample density forecasts.[17] The DSGE model includes seven observables and seven shocks; we separately evaluate the forecast densities for each of the target variables. We focus on the one-quarter-ahead forecast horizon.[18]

[15] Corradi and Swanson (2007) instead evaluate measures of in-sample fit of a benchmark DSGE model.

[16] See Section I in Smets and Wouters (2007) for a detailed description of the model.

[17] Smets and Wouters (2007) approximate the deciles of the predictive densities based on Gaussian kernel estimates, given the DSGE's assumption of normally distributed errors. We obtain the PITs using a linear interpolation for the inter-decile range.

[18] The sample period is from 1966:I to 2004:IV. The first one-quarter-ahead out-of-sample forecast is for 1985:I. From the 80 observations in each rolling window, 4 are used for pre-sampling: they are not included in

Table 5 reports the empirical results for the correct calibration of the DSGE density forecasts. The last two columns report the value of the κ_P^{CS} and C_P^{CS} tests that we propose in this paper. Asterisks '*' indicate rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4 with a HAC estimate of the covariance matrix. According to the critical values in Theorem 2, the density forecasts of inflation, hours and wages are well calibrated, although those of the interest rate, output growth, investment and consumption are not, according to at least one of our tests. When one allows for serial correlation under the null (that is, using the critical values implied by Theorem 4), there is more evidence of mis-specification in that the densities for all variables become poorly calibrated. Since the Ljung-Box test rejects that the PITs of all variables besides consumption are uncorrelated (in at least one of the moments – see the results reported in the same table), the version of our test that maintains serial correlation under the null is the appropriate one.

Figure 4 displays the cumulative distribution of the PITs for each the observables, together with critical values for correct calibration based on the κ_P^{CS} test in Theorem 2. The figures show that there are too few realizations of consumption and federal funds rate in the lowest quantiles of the distribution; that is, the DSGE underpredicts the target values. For the remaining variables, this test suggests proper calibration. For comparison, Figure 5 shows the estimated density forecasts of the observables, together with critical values based on Diebold et al. (1998). Diebold et al.'s (1998) methodology produces results similar to ours, except for output growth, hours worked and real wage forecasts, which are correctly specified according to our test and mis-specified according to Diebold et al.'s (1998) test.[19] Again, the most likely reason for the discrepancy appears to be the different nature of the test: our test is joint across deciles whereas the latter is pointwise.

Finally, Figure 6 depicts fan charts. The use of fan charts has been pioneered by the Bank of England to describe its best prediction of future inflation and its uncertainty to the general public.[20] A fan chart depicts ranges for likely values of the variable of interest

the likelihood. The total number of out-of-sample periods is 80. The model is estimated using Dynare codes by Smets and Wouters (2007). We create a sample of 150,000 draws for each rolling window estimation, discarding the first 20% of the draws. We use a step-size of 0.2 for the jumping distribution in the Metropolis-Hastings algorithm, resulting in rejection rates hovering around 0.4 across various estimation windows.

[19] Note that this is a fair comparison, since both Figures 4 and 5 are constructed under the maintained assumption of independence.

[20] The Bank of England's fan charts are obtained with a mix between statistical methods and judgement,

together with a line showing a measure of central tendency for the future outcomes (or point forecast). Typically, the possible values are plotted in different shades, where the darkest shade denotes more likely values than the lightest shade. Figure 6 depicts fan charts based on the Smets and Wouters (2007) model. The fan charts are based on parameter estimates from historical information up to 2000:IV: the solid line is the median of the one to four-quarter-ahead forecast distribution, whereas the shades depict the deciles of the forecast distribution. Interestingly, the forecasts are typically outside the most likely 80% of the forecast distribution at some horizon, especially for interest rates, hours and investment.

INSERT FIGURES 4, 5 AND 6 HERE

7 Conclusions

This paper proposes new tests for predictive density evaluation. They are designed to evaluate forecasting ability of the models in finite samples typically available to researchers. The techniques are based on Kolmogorov-Smirnov and Cramér-von Mises-type test statistics. We provide critical values of the tests for dynamically correctly specified models as well as for the case when the focus is on specific parts of the predictive density. We also propose methodologies that can be applied to dynamically mis-specified models and multiple-step-ahead forecast horizons. Empirical applications to the Survey of Professional Forecasters and a baseline DSGE model uncover that both SPF output growth and inflation density forecasts as well as DSGE-based forecasts of several macroeconomic aggregates are mis-specified.

References

[1] Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business and Economic Statistics* 25(2), 177-190.

[2] Ang, A., G. Bekaert and M. Wei (2007), "Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54, 1163-1212.

[3] Bache, I.W., A.S. Jore, J. Mitchell and S.P. Vahey (2011), "Combining VAR and DSGE Forecast Densities," *Journal of Economic Dynamics and Control* 35, 1659-1670.

and combining information from several models and risks in the economy. The fan charts that we report in this paper are instead based exclusively on the DSGE model we focus on, as our objective is to evaluate the DSGE model's predictive distribution.

[4] Bai, J. and S. Ng (2005), "Tests for Skewness, Kurtosis, and Normality for Time Series Data," *Journal of Business and Economic Statistics* 23(10), 49-60.

[5] Berkowitz, J. (2001), "Testing Density Forecasts, With Applications to Risk Management," *Journal of Business and Economic Statistics* 19(4), 465-474.

[6] Bank of England (2013), *Inflation Report Fan Charts*, May, http://www.bankofengland.co.uk/ publications/pages/inflationreport/irfanch.aspx

[7] Bank of Italy (2012), *Economic Bulletin* 67.

[8] Bontemps, C. and N. Meddahi (2012), "Testing Distributional Assumptions: A GMM Approach," *Journal of Applied Econometrics* 27(6), 978-1012.

[9] Box, G. and D. Pierce (1970), "Distribution of Residual Auto-correlation in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association* 65, 1509-1526.

[10] Brock, W. A., W. Dechert and J. Scheinkman (1987), "A Test for Independence based on the Correlation Dimension," *Working Paper*, University of Winconsin at Madison, University of Houston, and University of Chicago.

[11] Clements, M. P. (2004), "Evaluating the Bank of England Density Forecasts of Inflation," *The Economic Journal* 114, 844–866.

[12] Corradi, V. and N. R. Swanson (2006a), "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification," *Journal of Econometrics* 133, 779-806.

[13] Corradi, V. and N. R. Swanson (2006b), "Predictive Density Evaluation," In: G. Elliott, C. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting Vol. 1*, Elsevier, 197-284.

[14] Corradi, V. and N. R. Swanson (2006c), "Predictive density and conditional confidence interval accuracy tests," *Journal of Econometrics* 135(1–2), 187-228.

[15] Corradi, V., N. R. Swanson (2007), "Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data," *Journal of Econometrics* 136(2), 699-723.

[16] Christoffel, K., G. Coenen and A. Warne (2010), "Forecasting with DSGE Models," *ECB Working paper* 1185.

[17] Diks, C., V. Panchenkob and D. van Dijk (2011), "Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails," *Journal of Econometrics* 163, 215–230.

[18] Diebold, F. X., T. A. Gunther, and A. S. Tay (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review* 39(4), 863-883.

[19] Diebold F.X., A.S. Tay and K.F. Wallis (1999), "Evaluating Density Forecasts of Inflation: the Survey of Professional Forecasters." In: Engle R.F. and H. White, *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, Oxford University Press, 76-90.

[20] Edge, R. M. and R. S. Gürkaynak (2010), "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?" *Brookings Papers on Economic Activity* 41(2), 209-259.

[21] Edge, R. M., R. S. Gürkaynak, and B. Kısacıkoğlu (2013), "Judging the DSGE Model by Its Forecast," *mimeo.*

[22] Edge, R. M., M. T. Kiley and J. P. Laforte (2010), "A Comparison of Forecast Performance Between Federal Reserve Staff Forecasts, Simple Reduced-form Models, and a DSGE Model," *Journal of Applied Econometrics* 25(4), 720-754.

[23] Franses, P. H. and D. van Dijk (2003), "Selecting a Nonlinear Time Series Model using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics* 65, 727–744.

[24] González-Rivera, G. and Y. Sun (2013), "Generalized Autocontours: Evaluation of Multivariate Density Models," *mimeo.*

[25] Gneiting, T. and A. E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association* 102, 359-378.

[26] Gürkaynak, R. S., B. Kisacikoglu and B. Rossi (2013), "Do DSGE Models Forecast More Accurately Out-of-Sample than VAR Models?" In: T. Fomby, L. Kilian and A. Murphy (eds.), *Advances in Econometrics: VAR Models in Macroeconomics –New Developments and Applications Vol. 31*, forthcoming.

[27] Hong, Y. M. and H. Li (2005), "Nonparametric Specification Testing for Continuous Time Models with Applications to Term Structure of Interest Rates," *Review of Financial Studies* 18(1), 37-84.

[28] Hong, Y., Li, H. and F. Zhao (2007), "Can the Random Walk Model Be Beaten in Out-of-sample Density Forecasts? Evidence From Intraday Foreign Exchange Rates," *Journal of Econometrics* 141(2), 736–776.

[29] Inoue, A. (2001), "Testing for Distributional Change in Time Series," *Econometric Theory* 17, 156-187.

[30] Newey, W.K. and K.D. West (1987), "A Simple, Positive semi-definite, Heteroskedasticityand Auto-correlation Consistent Covariance Matrix," *Econometrica* 55(3), 703-708.

[31] Persson, J. (1974), "Comments on Estimations and Tests of EEG Amplitude Distributions," *Electroencephalography and Clinical Neurophysiology* 37, 309-313.

[32] Shorack, G. R. and J. A. Wellner (1986), *Empirical Processes with Applications to Statistics*, Wiley.

[33] Smets, F. and R. Wouters (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review* 97(3), 586-607.

[34] Wolters, M. H. (2012), "Evaluating Point and Density Forecasts of DSGE Models," *MPRA Working Paper* 36147.

[35] Weiss, M. S. (1973), "Modifications of the Kolmogorov-Smirnov Statistic for Use with Correlated Data," *Journal of the American Statistical Association* 74, 872-875.

8 Appendix A. Proofs

The appendix provides the proofs for Theorems 1, 2, 3 and 4.

Proof of Theorem 1. (i) The true joint conditional predictive density of $\{y_{t+1}\}_{t=R}^{T}$ can be decomposed as $\phi_0(y_{T+1},...,y_R|\mathcal{F}_R) = \phi_0(y_{T+1}|\mathcal{F}_T)\,\phi_0(y_T|\mathcal{F}_{T-1})...\phi_0(y_{R+1}|\mathcal{F}_R)$, where R is finite by Assumption 1(iii) (which guarantees that we condition on a finite information set). Let $q(z_{R+1},...,z_{T+1}|\mathcal{F}_R)$ denote the conditional joint density of the probability integral transforms. Then, given that $y_{t+1} = \Phi_{t+1}^{-1}(z_{t+1}|\Im_t)$, where $\Im_t \subseteq \mathcal{F}_t$, we can re-write the joint density of the PITs as $q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = \phi_0\left(\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)|\mathcal{F}_R\right)...\phi_0\left(\Phi_T^{-1}(z_T|\Im_{T-1})|\mathcal{F}_{T-1}\right) \times ... \times \phi_0\left(\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)|\mathcal{F}_T\right)$.

By using the change of variables formula,

$$
\begin{aligned}
q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = & \begin{vmatrix} \left(\partial\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)/\partial z_{R+1}\right) & ... & \left(\partial\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)/\partial z_{T+1}\right) \\ ... & ... & ... \\ \left(\partial\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)/\partial z_{R+1}\right) & ... & \left(\partial\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)/\partial z_{T+1}\right) \end{vmatrix} \\
& \times \phi_0\left(\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)|\mathcal{F}_R\right)...\phi_0\left(\Phi_T^{-1}(z_T|\Im_{T-1})|\mathcal{F}_{T-1}\right)\phi_0\left(\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)|\mathcal{F}_T\right) \\
& = \left(1/\phi_{R+1}(y_{R+1}|\Im_R)\right)...\left(1/\phi_T(y_T|\Im_{T-1})\right)\left(1/\phi_{T+1}(y_{T+1}|\Im_T)\right) \times \\
& \times \phi_0(y_{R+1}|\mathcal{F}_R)...\phi_0(y_T|\mathcal{F}_{T-1})\,\phi_0(y_{T+1}|\mathcal{F}_T),
\end{aligned}
$$

where the last equality holds because the Jacobian is lower triangular provided we are in a conditional forecasting framework and thus $\{y_{t+1,},...,y_{T+1}\} \notin \Im_t$ at any time t. Then,

$$
q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = \frac{\phi_0(y_{R+1}|\mathcal{F}_R)}{\phi_{R+1}(y_{R+1}|\Im_R)} \times ... \times \frac{\phi_0(y_T|\mathcal{F}_{T-1})}{\phi_T(y_T|\Im_{T-1})} \times \frac{\phi_0(y_{T+1}|\mathcal{F}_T)}{\phi_{T+1}(y_{T+1}|\Im_T)}.
$$

Now suppose that $\Im_t = \Im_{t-R+1}^t$, where $\Im_{t-R+1}^t$ contains only data available from time $t-R+1$ to time t. In other words, $\Im_{t-R+1}^t$ differs from $\Im_t$ because the rolling window does not use all the available information in the sample. If $\phi_{t+1}(y_{t+1}|\Im_t) = \phi_{t+1}\left(y_{t+1}|\Im_{t-R+1}^t\right) = \phi_{t+1}(y_{t+1}|\mathcal{F}_t)$ (the condition imposed by Assumption 2), i.e. when $\Im_{t-R+1}^t$ contains all relevant past information, then we could re-write the above as

$$
q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = \frac{\phi_0(y_{R+1}|\mathcal{F}_R)}{\phi_{R+1}(y_{R+1}|\Im_1^R)} \times ... \times \frac{\phi_0(y_T|\mathcal{F}_{T-1})}{\phi_T\left(y_T|\Im_{T-R}^{T-1}\right)} \times \frac{\phi_0(y_{T+1}|\mathcal{F}_T)}{\phi_{T+1}\left(y_{T+1}|\Im_{T-R+1}^T\right)}
$$

It follows that, under the null, each ratio yields a $U(0,1)$ variable (since the pdf is the unit line), thus the joint distribution is a multivariate $U(0,1)$. In addition, since the joint distribution is the product of the marginals, then $\{z_{t+1}\}_{t=R}^{T}$ is *iid* $U(0,1)$.

(ii) Under H_0, z_{t+1} is uniformly distributed on $[0,1]$. Then, from Proposition 1 in Shorack and Wellner (1986, p. 131), $\Psi_P(\pi, r) = P^{-1/2}\sum_{t=R}^{T}\left(1\left\{\Phi_{t+1}\left(y_{t+1}|\Im_{t-R+1}^{t}\right) \leq r\right\} - r\right)$ weakly converges (considered as variables in the space $[0,1]^2 \times \mathbb{R}$) to the normal $\Psi^{\circ}(.,.)$, with mean zero and auto-covariance function $E\left[\Psi^{\circ}(r_1)\Psi^{\circ}(r_2)\right] = \left[\inf(r_1, r_2) - r_1 r_2\right]$. ■

Proof of Theorem 2. The theorem follows from Theorem 1 by the Continuous Mapping theorem. ■

Proof of Corollary 3. The theorem follows directly from part (i) in Theorem 1 and Berkowitz (2001). ■

Proof of Theorem 4. Theorem 4 follows from Inoue (2001) by letting (in the referenced paper's notation) $r = 1$.

Asymptotic Distribution of the Continuous Rank Probability Score, $\mathbf{CRPS}_T$. Let the (rescaled average) Continuous Rank Probability Score (CRPS, Gneiting and Raftery, 2007) be defined as:

$$CRPS_P = -\int_{\mathbb{R}}\left\{P^{-1/2}\sum_{t=R}^{T}\left[\Phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) - 1\left(y_{t+h} \leq y\right)\right]\right\}^2 dy$$

Let the null hypothesis be $\widetilde{H}_0 : \Phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) - 1\left(y \geq y_{t+h}\right) = 0$. Under Assumptions 1(ii), 1(iii) and 3, where 3(ii) is replaced by Assumption 3(ii)*:

Assumption 3(ii):* $\Pr\left(\Phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) \leq r_1, \Phi_{t+h+d}\left(y|\Im_{t-R+1}^{t}\right) \leq r_2\right) = \widetilde{F}_d(r_1, r_2)$,

$\Pr\left(\Phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) \leq r\right) = \widetilde{F}(r)$, *where* $\widetilde{F}(.)$ *is continuous and* $\widetilde{F}_d(.,.)$ *and* $\widetilde{F}(.)$ *are the distribution functions of the random variable* y.

It follows from Theorems 2.1 and 2.2 in Inoue (2001, letting $r = 1$ in his notation) that

$$\int_{\mathbb{R}}\left\{P^{-1/2}\sum_{t=R}^{T}\left[\Phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) - 1\left(y_{t+h} \leq y\right)\right]\right\}^2 dy \equiv T_2 \Rightarrow \int_{\mathbb{R}}\widetilde{K}(y)^2 dy, \tag{9}$$

where $\widetilde{K}(.)$ is a mean-zero Gaussian process with covariance kernel $E\left(\widetilde{K}(y_1)\widetilde{K}(y_2)\right) = \widetilde{\sigma}(y_1, y_2)$, where $\widetilde{\sigma}(y_1, y_2) = \sum_{d=-\infty}^{\infty}\left[\widetilde{F}_d(y_1, y_2) - \widetilde{F}(y_1)\widetilde{F}(y_2)\right]$. ■

9 Appendix B. Tables and Figures

Table 1. Critical Values

		$\kappa^{CS}_{\alpha;P}$			$C^{CS}_{\alpha;P}$		
	$\alpha:$	0.01	0.05	0.10	0.01	0.05	0.10
Panel A. Tests on the Whole Distribution							
Correct Specification Test		2.25	1.51	1.19	0.74	0.46	0.35
Panel B. Tests on Specific Parts of the Distribution							
Right Tail	$r \in (0, 0.25]$	1.16	0.70	0.52	0.50	0.30	0.22
Right Half	$r \in (0, 0.50]$	1.96	1.25	0.98	0.81	0.49	0.36
Left Half	$r \in [0.50, 1)$	2.04	1.31	1.01	0.91	0.55	0.40
Left Tail	$r \in [0.75, 1)$	1.34	0.82	0.61	0.66	0.40	0.29
Center	$r \in [0.25, 0.75]$	2.21	1.48	1.16	1.13	0.69	0.50
Tails	$r \in \{(0, 0.25] \cup [0.75, 1)\}$	1.45	0.95	0.74	0.43	0.28	0.22

Note: Panel A reports critical values for the test statistics κ^{CS}_P and C^{CS}_P at the 1%,5% and 10% nominal sizes ($\alpha = 0.01, 0.05$ and 0.10). Panel B reports critical values for the same statistics for specific parts of the distributions, indicated in the second column. The number of Monte Carlo replications is 1,000,000. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.05.

Table 2: Size Properties

DGP S1 (IID Case)							
		κ_P^{CS}			C_P^{CS}		
P	$R:$	50	100	200	50	100	200
100		0.05	0.06	0.06	0.05	0.06	0.05
200		0.05	0.05	0.05	0.05	0.05	0.05
500		0.05	0.05	0.05	0.05	0.05	0.05
1000		0.05	0.05	0.05	0.05	0.05	0.05
DGP S2 (IID Case)							
		κ_P^{CS}			C_P^{CS}		
P	$R:$	50	100	200	50	100	200
100		0.06	0.06	0.05	0.05	0.05	0.05
200		0.05	0.05	0.05	0.05	0.05	0.05
500		0.05	0.06	0.05	0.05	0.06	0.05
1000		0.05	0.05	0.05	0.05	0.06	0.05
DGP S3 (Serially Correlated Case)							
		κ_P^{CS}			C_P^{CS}		
P	$R:$	50	100	200	50	100	200
100		0.07	0.12	0.11	0.04	0.07	0.06
200		0.09	0.13	0.09	0.05	0.07	0.05
500		0.09	0.11	0.08	0.05	0.06	0.05
1000		0.09	0.10	0.10	0.05	0.05	0.06

Note: The table reports empirical rejection frequencies for the test statistics κ_P^{CS} and C_P^{CS} in eqs. (4) and (5) at the 5% nominal size for various values of P and R. The number of Monte Carlo replications is 5,000. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.05. Critical values for DGP S1 and DGP S2 are those reported in Table 1, Panel A. For DGP S3, the critical values are simulated with a HAC estimate of a covariance matrix.

Table 3. Power Properties

DGP P		
c	κ_P^{CS}	C_P^{CS}
0	0.05	0.05
0.10	0.35	0.40
0.15	0.80	0.91
0.20	0.99	1.00
0.25	1.00	1.00

Note: The table reports empirical rejection frequencies for the test statistics κ_P^{CS} and C_P^{CS} in eqs. (4) and (5) at the 5% nominal size for P=960 and R=40. The number of Monte Carlo replications is 5,000. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.05. Critical values for DGP S1 and DGP S2 are those reported in Table 1, Panel A.

Table 4: Correct Specification Tests for SPF's Probability Forecast Distribution

Series Name:	GDP Growth		GDP Deflator Growth	
Forecast Horizon (in rows):	Correct Specification Tests			
	κ_P^{CS}	C_P^{CS}	κ_P^{CS}	C_P^{CS}
0	2.10*†	0.81*†	15.71*†	5.12*†
1	0.59†	0.12†	23.40*†	10.16*†

Note: Asterisks '*' indicate rejection at 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4. The latter are simulated conditional on the data (i.e. conditional on the variance-covariance matrix for the PIT). The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.05.

Table 5: Baseline DSGE Distribution of Forecasts (1985:I - 2004:IV)

Variable	LB: $(z_t - \bar{z})$	LB: $(z_t - \bar{z})^2$	κ_P^{CS}	C_P^{CS}
Consumption (real)	0.09	0.61	4.42 *†	1.84 *†
Investment (real)	0.00 *	0.52	0.34 †	0.16 *†
Output Growth (real)	0.00 *	0.28	1.46 †	0.66 *†
Inflation	0.07 *	0.83	1.20 †	0.45 †
Hours	0.00 *	0.53	0.65 †	0.31 †
Wages (real)	0.82	0.04 *	1.06 †	0.27 †
Federal Funds Rate	0.00 *	0.00 *	3.12 *†	1.37 *†

Note: The column labeled "LB" indicates p-values of the Ljung-Box test statistic for absence of serial correlation; values marked by '*' indicate rejections at 5% significance level. For the κ_P^{CS} and C_P^{CS} tests, '*' indicates rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.05.

Figure 1. Representative Fan Charts from the DSGE model in 2000:IV

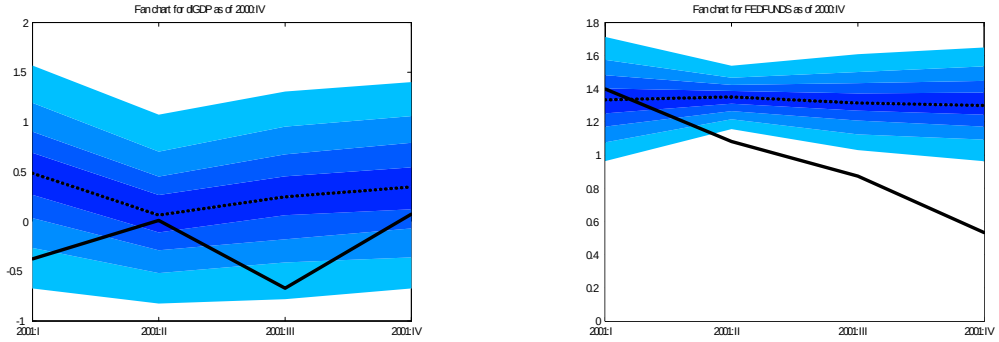
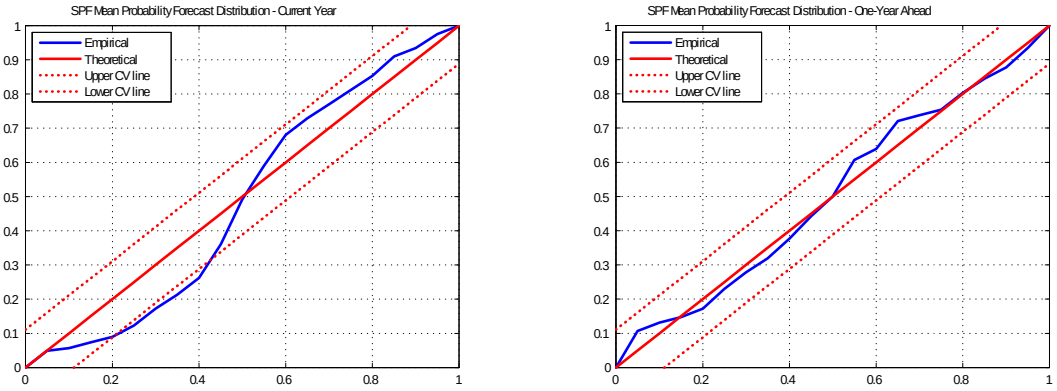
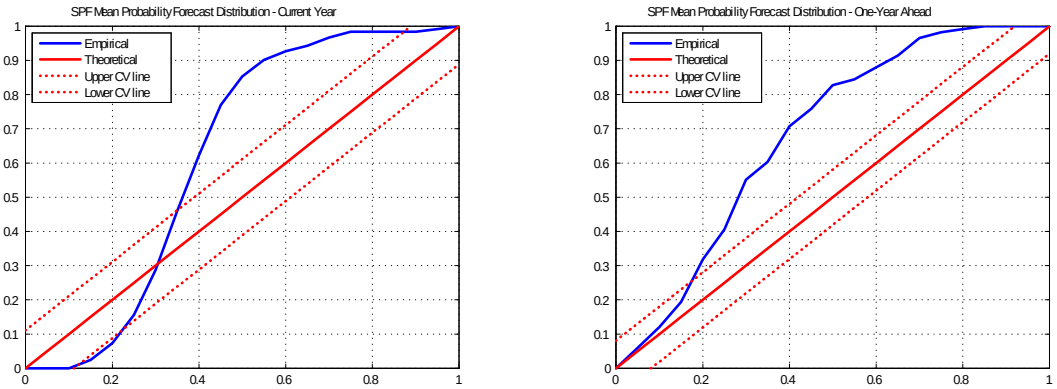


Figure 2. SPF Mean Probability Forecast Distribution

Panel A: GDP Growth (1981:III-2011:IV)



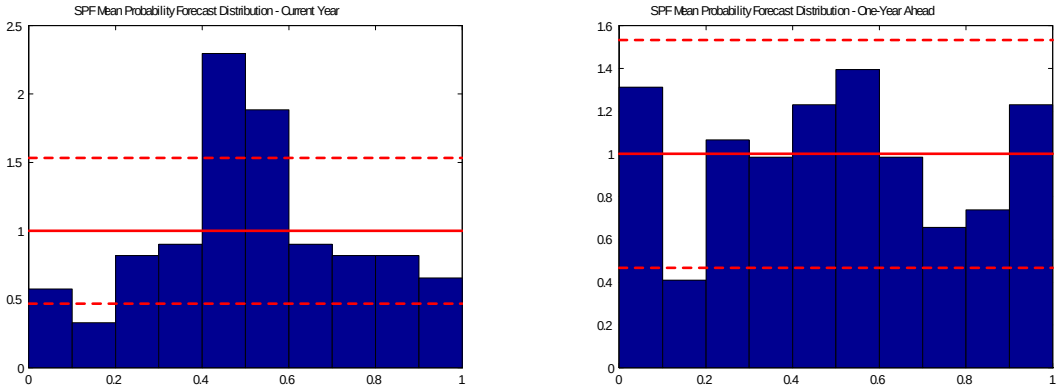
Panel B: GDP Deflator Growth (1981:III-2009:IV)



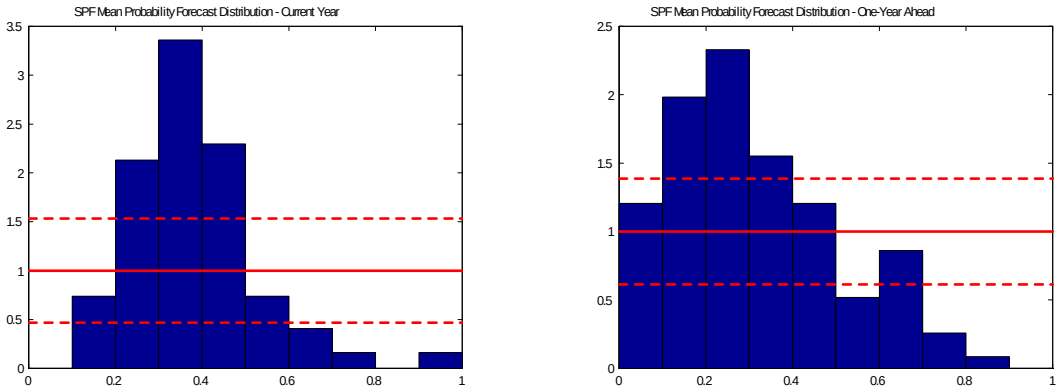
Note: The figure shows the empirical CDF of the PITs (solid line), the CDF of the PITs under the null hypothesis (the 45 degree line) and the critical values based on the κ_P^{CS} test.

Figure 3. SPF Mean Probability Forecast Distribution

Panel A: GDP Growth (1981:III-2011:IV)

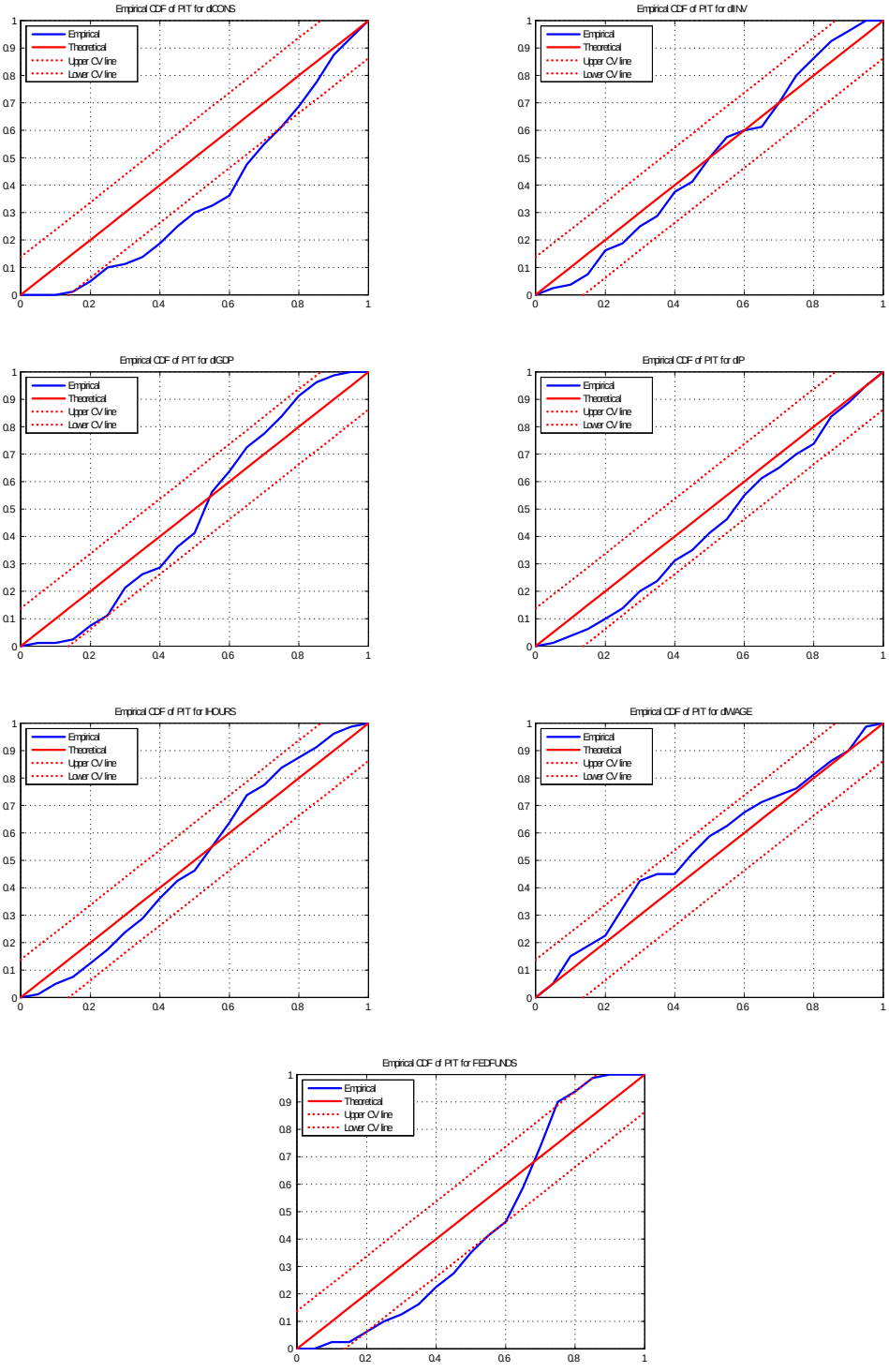


Panel B: GDP Deflator Growth (1981:III-2009:IV)



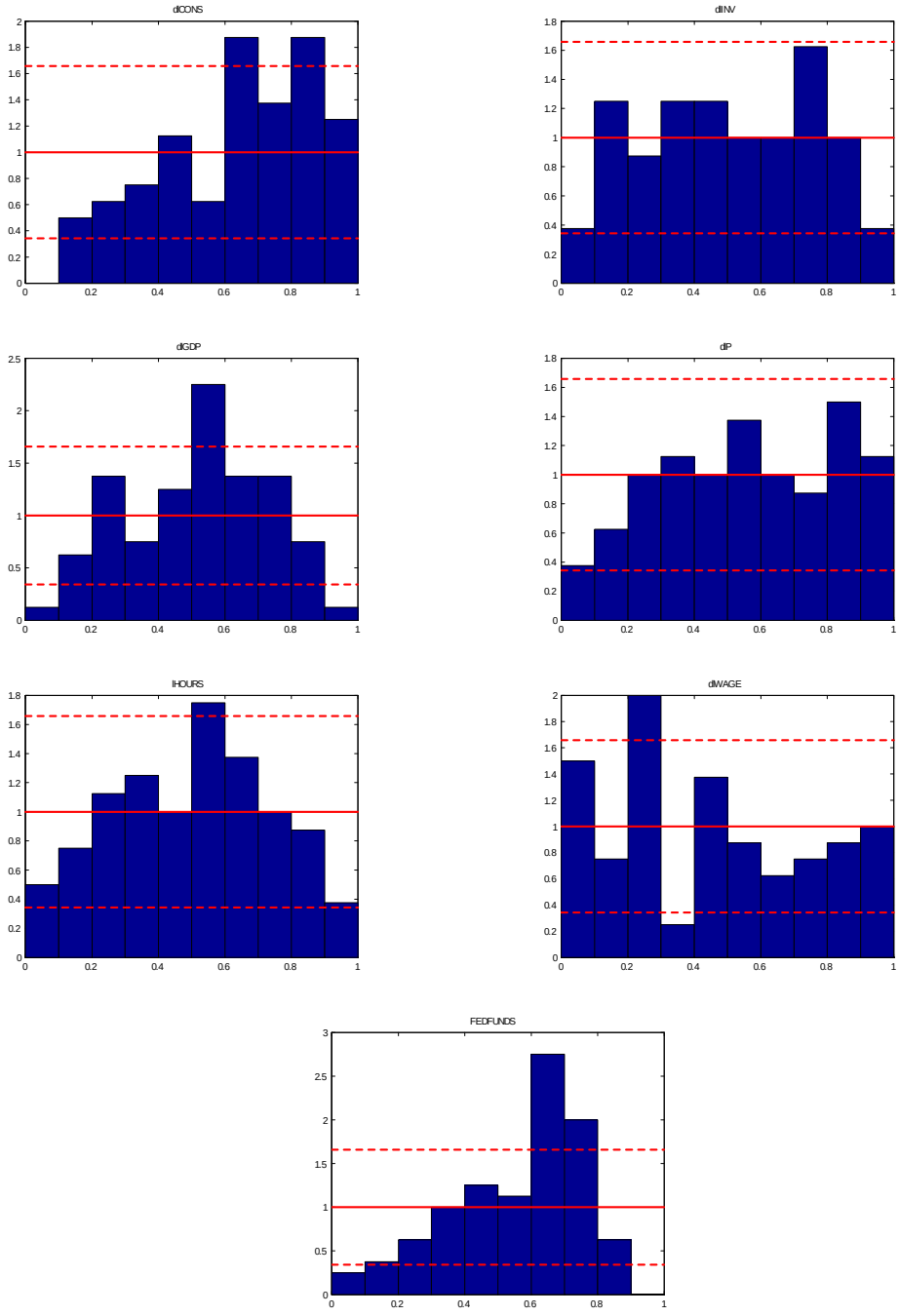
Note: The figures show the pdf of the PITs (normalized) and the 95% critical values approximated under Diebold et al.'s (1998) binomial distribution (dashed lines), constructed using a normal approximation.

Figure 4. Baseline DSGE Distribution of Forecasts (1985:I-2004:IV)



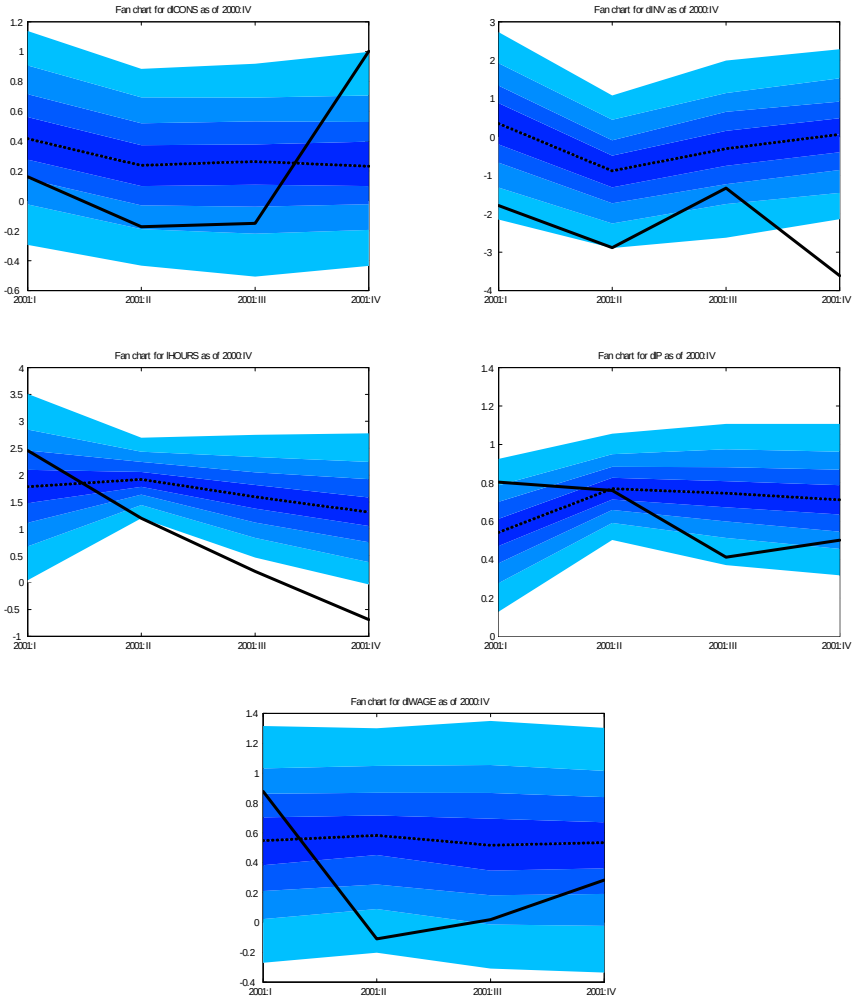
Note: The figures show the empirical CDF of the PITs (solid line), the CDF of the PITs under the null hypothesis (the 45 degree line) and the 95% confidence bands based on critical values of κ_P^{CS} test reported in Table 1, Panel A. Results are based on a rolling window of size $R = 80$.

Figure 5. Baseline DSGE Distribution of Forecasts (1985:I-2004:IV)



Note: The figures show the pdf of the PITs (normalized) and the 95% critical values approximated under Diebold et al.'s (1998) binomial distribution (dashed lines), constructed using a normal approximation. The results are based on a rolling window of size $R = 80$.

Figure 6. Baseline DSGE Fan Charts as of 2000:IV



Note: The figure shows fan charts obtained by estimating the baseline DSGE with data up to 2000:IV, prior to 2001:I-2001:IV recession. Depicted are the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th deciles of the predictive distribution for one to four-quarter-ahead out-of-sample forecasts. The solid line represents the actual realizations of the data, while the dotted lines represent the median forecast.